Точные оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированных методов обучения

Фрей Александр Ильич

Московский физико-технический институт
(Государственный университет)
Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы» (ВЦ РАН)

Научный руководитель: д.ф.-м.н. Воронцов Константин Вячеславович

16 июня 2010

Проблема переобучения

- Строки таблицы $\{x_1 \ldots x_\ell, x_{\ell+1}, x_L\}$ — объекты полной выборки
- Столбцы $\{a_1 \ldots a_D\}$ — векторы ошибок алгоритмов

	a_1	a_2	$\ldots$	a_d	$\ldots$	a_D
x_1	0	1	$\ldots$	**0**	$\ldots$	1
$\ldots$	1	1	$\ldots$	**0**	$\ldots$	0
x_ℓ	0	0	$\ldots$	**0**	$\ldots$	0
$x_{\ell+1}$	1	1	$\ldots$	**1**	$\ldots$	1
$\ldots$	1	0	$\ldots$	**1**	$\ldots$	0
x_L	0	0	$\ldots$	**1**	$\ldots$	0

- Метод обучения — минимизация эмпирического риска
- Цель: получить точные, вычислительно-эффективные оценки вероятности переобучения.

Методы вывода формул для вероятности переобучения

1. Метод производящих и запрещающих объектов
   - Монотонная цепочка и сетка
2. Блочная оценка
   - Пара алгоритмов
3. Рекуррентное вычисление вероятности переобучения по заданной матрице ошибок
   - Теоретический инструмент для доказательства универсальных оценок
4. Гипотеза t-слоев и метод β-многочленов
   - Точные оценки для унимодальных цепочек
   - Приближенные оценки для унимодальных сеток

Методы вывода формул для вероятности переобучения

1. Метод производящих и запрещающих объектов
   - Монотонная цепочка и сетка
2. Блочная оценка
   - Пара алгоритмов
3. Рекуррентное вычисление вероятности переобучения по заданной матрице ошибок
   - Теоретический инструмент для доказательства универсальных оценок
4. Гипотеза t-слоев и метод β-многочленов
   - Точные оценки для унимодальных цепочек
   - Приближенные оценки для унимодальных сеток
5. Метод разбиения множества алгоритмов на орбиты
   - Пучок монотонных цепочек
   - Полный слой, полный куб алгоритмов
   - Шар алгоритмов
   - Точные оценки для монотонных и унимодальных сеток

Рандомизированный метод обучения

- Генеральная выборка $\mathbb{X} = \left(x_i\right)_{i=1}^{L}$
- Алгоритм — бинарный вектор $a \equiv \left(a(x_i)\right)_{i=1}^{L}$ длины L
- Конечное множество алгоритмов $A = \{a_1, \ldots, a_D\}$
- Метод обучения:

$$\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^{\ell} \to A \quad — \quad \text{детерминированный;}$$
$$\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^{\ell} \to \{f : A \to [0,1]\} \quad — \quad \text{рандомизированный;}$$

- *Вклад* алгоритма $a \in A$ в вероятность переобучения:

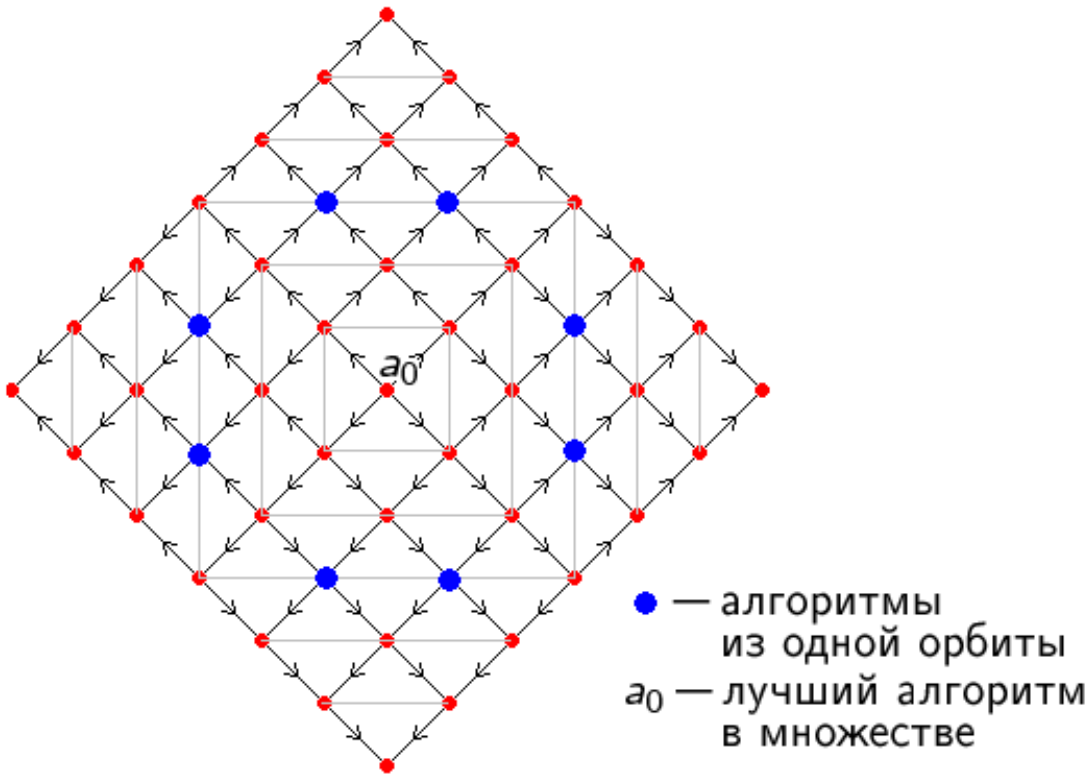
$$Q_\mu(\varepsilon, a, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \mu(A, X, a) \left[\delta(a, X) \geq \varepsilon\right].$$

- *Вероятность переобучения*: $Q_\mu(\varepsilon, A) = \sum_{a \in A} Q_\mu(\varepsilon, a, A)$.

Орбита алгоритмов

Граф смежности множества алгоритмов:

- Вершины соответствуют алгоритмам
- Ребро $a_1 \longrightarrow a_2$:
  - соединяет алгоритмы, различающиеся на одном объекте
  - идет в направлении возрастания числа ошибок



Теорема о равном вкладе алгоритмов одной орбиты

- Группа перестановок объектов выборки S_L *действует* на множестве всех алгоритмов перестановками координат.
- *Группой симметрий* $\mathrm{Sym}(A) \subset S_L$ множества алгоритмов назовем стационарную подгруппу S_L:

$$\mathrm{Sym}(A) = \{\pi \in S_L : \pi(A) = A\}.$$

- Пусть $\pi \in \mathrm{Sym}(A)$, $a \in A$ — алгоритм множества A. Тогда a и πa дают **равный вклад** в вероятность переобучения.

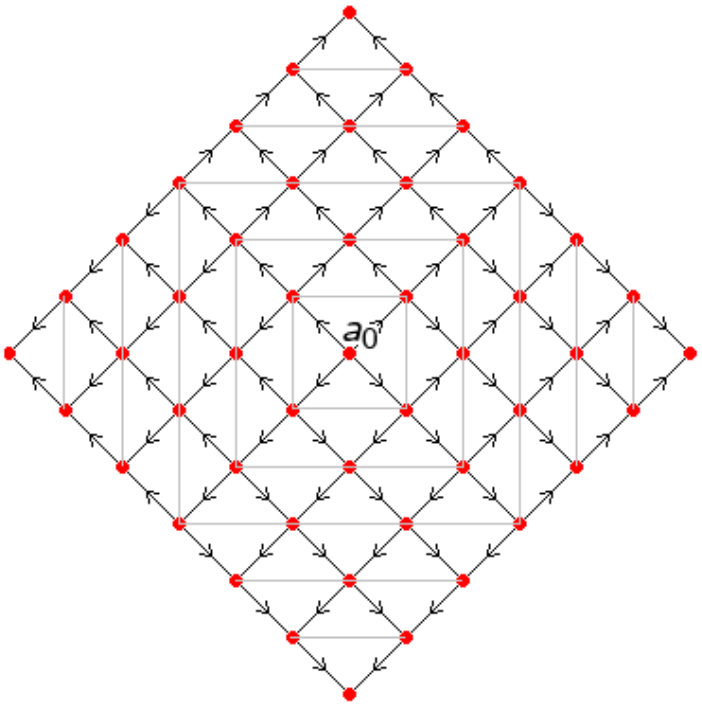
Теорема

Вероятность переобучения метода μ записывается в виде:

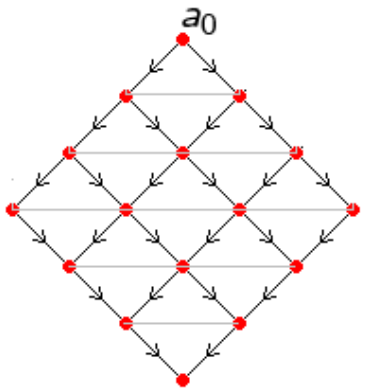
$$Q_\mu(\varepsilon, A) = \frac{1}{C_L^\ell} \sum_{\omega \in \Omega(A)} |\omega| \sum_{X \in [\mathbb{X}]^\ell} \mu(A, X, a) \left[\delta(a_\omega, X) \geq \varepsilon\right]. \qquad (1)$$

Модельные семейства алгоритмов

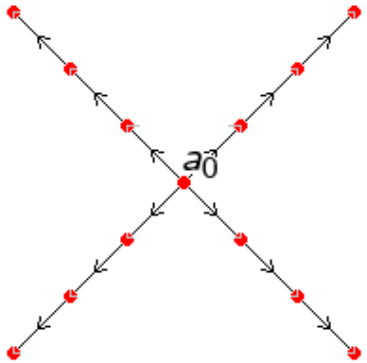
Унимодальная сетка размерности h является реалистичной моделью *связного* параметрического семейства алгоритмов.



Унимодальная сетка



Монотонная сетка



Связка монотонных цепочек

Точные формулы для вероятности переобучения

Теорема (Связка из p монотонных цепочек)

$$Q_\mu(\varepsilon, A) = \sum_{h=0}^{D} \sum_{S=h}^{pD} \sum_{F=0}^{p} \frac{|\omega_h| R_{D,p}^h(S,F)}{1+S} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}(s(\varepsilon)),$$

где $L' = L - S - F$, $\ell' = \ell - F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m + h - \varepsilon k) \right\rfloor$;
$|\omega_h| = 1$ *при* $h = 0$ *и* $|\omega_h| = p$ *при* $h \geq 1$;
$H_{L'}^{\ell',m}(s)$ — *функция гипергеометрического распределения;*
$R_{D,p}^h(S,F)$ — *число способов представить* S *в виде суммы* p *неотрицательных слагаемых* $S = t_1 + \ldots + t_p$, *каждое из которых не превосходит* D, *некоторые* F *слагаемых строго меньше* D, *а* $t_1 \geq h$.

Точные формулы для вероятности переобучения

- Множество орбит как монотонных, так и унимодальных сеток, индексированно диаграммами Юнга $Y_*^{h,D}$.

Теорема ($Q_\mu(\varepsilon, A)$ для h-мерной монотонной сетки)

$$Q_\mu(\varepsilon, A) = \sum_{\lambda \in Y_*^{h,D}} \sum_{\substack{\vec{t} \geq \lambda, \\ \|\vec{t}\| \leq D}} \frac{|S_h \lambda|}{\prod_j (t_j + 1)} \frac{C_{L'}^{\ell'}}{C_L^{\ell}} H_{L'}^{\ell', m}(s_0),$$

Теорема ($Q_\mu(\varepsilon, A)$ для h-мерной унимодальной сетки)

$$Q_\mu(\varepsilon, A) = \sum_{\lambda \in Y_*^{h,D}} \sum_{\substack{\vec{t} \geq \lambda, \\ \|\vec{t}\| \leq D}} \sum_{\substack{\vec{t'} \geq 0, \\ \|\vec{t'}\| \leq D}} \frac{|S_h \lambda| \cdot 2^{|\lambda > 0|}}{\prod_j (t_j + t'_j + 1)} \frac{C_{L'}^{\ell'}}{C_L^{\ell}} H_{L'}^{\ell', m}(s_0),$$

Сравнение сеток и связки монотонных цепочек

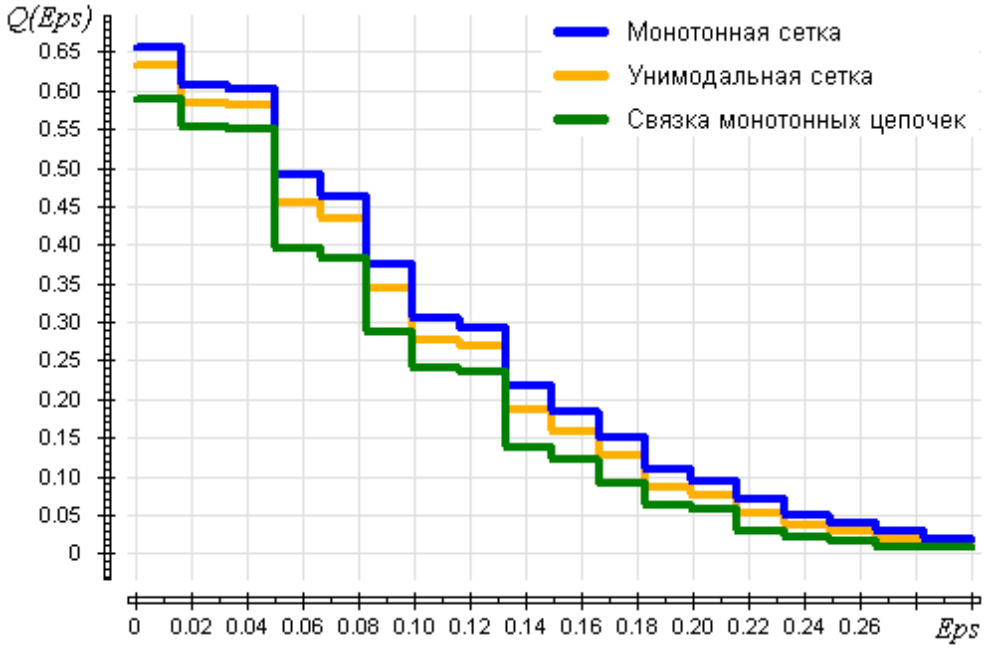
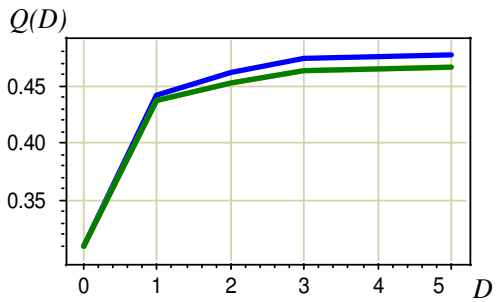
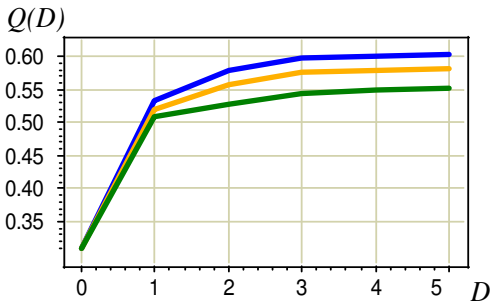


Рис.: Сравнение при разных ε. $D = 5$, $m = 5$, $L = 50$, $\ell = 30$.





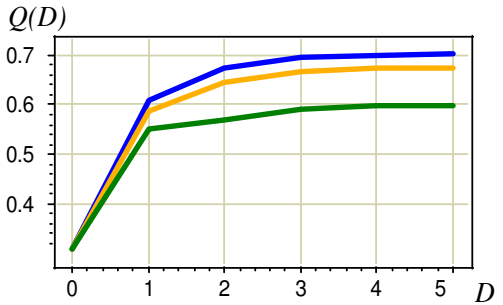


Рис.: Сравнение при разных D, в размерностях $H = 1(2)$, $H = 2(4)$ и $H = 4(6)$. $\varepsilon = 0.04$, $m = 5$, $L = 50$, $\ell = 30$.

Точные формулы для вероятности переобучения

Теорема (Полный слой A_m алгоритмов с m ошибками)

$$Q_\mu(\varepsilon, A) = \left[\varepsilon k \le m \le L - \varepsilon\ell\right].$$

Теорема (Полный куб алгоритмов $A = \{0,1\}^L$)

$$Q_\mu(\varepsilon, A) = \frac{1}{2^k} \sum_{m=\lceil \varepsilon k \rceil}^{k} C_k^m.$$

Теорема (Сечение $B_r(a_0) \cap A_m$ шара центральным слоем)

$$Q_\mu(\varepsilon, A) = H_L^{\ell,m}(s_d(\varepsilon) + \lfloor r/2 \rfloor),$$

где $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$,
$H_L^{\ell,m}(s)$ *— функция гипергеометрического распределения.*

Точные формулы для вероятности переобучения

Теорема (Шар алгоритмов)

Пусть $A = B_r(a_0)$ *— шар алгоритмов,* $m = n(a_0, \mathbb{X})$, $r \leq \min(m, L - m)$. *Тогда вероятность переобучения рандомизированного метода минимизации эмпирического риска записывается в виде*

$$Q_\mu(\varepsilon, A) = \sum_{i=0}^{r} h_L^{\ell,m}(i) \frac{\sum\limits_{p=0}^{r-i} \sum\limits_{q=q_0}^{r-i-p} C_{m-i}^p C_{k-(m-i)}^q}{\sum\limits_{p=0}^{r-i} C_k^p} + \sum_{i=r+1}^{\lfloor s'_d(\varepsilon) \rfloor} h_L^{\ell,m}(i),$$

где $q_0 = \max(\lceil \varepsilon k + i + p - m \rceil, 0)$, $s'_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k) + \frac{rk}{L}$,
$h_L^{\ell,m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}$.

Результаты, выносимые на защиту

1. Предложен рандомизированный метод минимизации эмпирического риска;
2. Разработан новый теоретико-групповой метод вывода оценок для вероятности переобучения;
3. Получены точные оценки вероятности переобучения для:
   - Пучка монотонных цепочек;
   - Полного слоя, полного куба алгоритмов;
   - Монотонных и унимодальных сеток;
4. Экспериментально показано, что в широком диапазоне параметров вероятность переобучения связки монотонных цепочек не превосходит вероятности переобучения многомерных монотонных и унимодальных сеток.